AE 22: Multiple regression pt 2

YOUR NAME GOES HERE

2021-11-11

```
library(tidyverse)
library(tidymodels)
library(scatterplot3d)
library(viridis)
```

Learning goals

By the end of today, you will be able to…

- use and understand the R^2 statistic
- model interactions between variables
- conduct a hypothesis test about a particular β_i

To begin, we'll work with a dataset on iris flowers contained with base `R`. This dataset comes from an old statisical analysis paper aptly titled *The use of multiple measurements in taxonomic problems*

```
data(iris)
glimpse(iris)
```

```
## Rows: 150
## Columns: 5
## $ Sepal.Length <dbl> 5.1, 4.9, 4.7, 4.6, 5.0, 5.4, 4.6, 5.0, 4.4, 4.9, 5.4, 4.~
## $ Sepal.Width  <dbl> 3.5, 3.0, 3.2, 3.1, 3.6, 3.9, 3.4, 3.4, 2.9, 3.1, 3.7, 3.~
## $ Petal.Length <dbl> 1.4, 1.4, 1.3, 1.5, 1.4, 1.7, 1.4, 1.5, 1.4, 1.5, 1.5, 1.~
## $ Petal.Width  <dbl> 0.2, 0.2, 0.2, 0.2, 0.2, 0.4, 0.3, 0.2, 0.2, 0.1, 0.2, 0.~
## $ Species      <fct> setosa, setosa, setosa, setosa, setosa, setosa, setosa, s~
```

See `?iris` for more information.

R^2 discussion

Exercise 1

- In your own words, what is R^2?

[write here]

Exercise 2

Conceptualize R^2

$$R^2 = 1 - \frac{\text{sum of squared error}}{\text{sum of square distance from mean in data}}$$

Let's focus on the second term to build intuition.

- The numerator "sum of squared error" is the amount of variability *not* explained by the model.
- The denominator is proportional to the variance in the data, i.e. the amount of variability in the data.
- Together, this second term represents the proportion of variability *not* explained by the model.

If the proportion *not* explained is 0, the model explains all variability and $R^2 = 1 - 0 = 1$.

If the proportion *not* explained is 1, i.e. the model does not explain any variability, then $R^2 = 1 - 1 = 0$.

- Revisit your explanation in exercise 1, would you update your description in any way? If so, do so here:

[updated explanation]

Example: R^2

Suppose that we want to go out and collect more data on irises, but measuring several parts of each iris flower is time consuming. To save on the time it takes to collect future data, let's see if one of the observations (petal length) could be predicted from just measuring a flower's sepal. (This might make sense since the sepal is the part of the flower below the bud that contains the petals before it blooms)

To do this, we will setup three linear models:

1. predict petal length from sepal length
2. predict petal length from sepal width
3. predict petal length from both sepal width and sepal length

Exercise 3.

- Why is it useful to investigate each of these models?

We investigate each model so that we can compare their fit and figure out the best combination of sepal measurement predictors.

- Before writing any code, how do you think R^2 will vary between models 1, 2 and 3? Why?

Model 3 contains multiple predictors, so it should have the highest R^2.

Exercise 4

- Fit each linear regression model above and compare R^2.

```r
# code here
fit1 = linear_reg() %>%
  set_engine("lm") %>%
  fit(Petal.Length ~ Sepal.Length, data = iris)

fit2 = linear_reg() %>%
  set_engine("lm") %>%
  fit(Petal.Length ~ Sepal.Width, data = iris)

fit3 = linear_reg() %>%
  set_engine("lm") %>%
  fit(Petal.Length ~ Sepal.Length + Sepal.Width, data = iris)

glance(fit1)$r.squared
```

```
## [1] 0.7599546
```

```
glance(fit2)$r.squared
```

```
## [1] 0.1835609
```

```
glance(fit3)$r.squared
```

```
## [1] 0.867686
```

- Do your results match your expectations? Which model offers the best fit? Why?

The third model indeed has the highest R^2 since it has the most explanatory variables.

- Now compare 'adjusted' R^2 between models. Which model offers the best fit? Why? See here for details on finding adjusted R^2.

```
# code here
glance(fit1)$adj.r.squared
```

```
## [1] 0.7583327
```

```
glance(fit2)$adj.r.squared
```

```
## [1] 0.1780444
```

```
glance(fit3)$adj.r.squared
```

```
## [1] 0.8658858
```

Model 3 continues to offer the best fit here.

Hypothesis testing in a regression framework

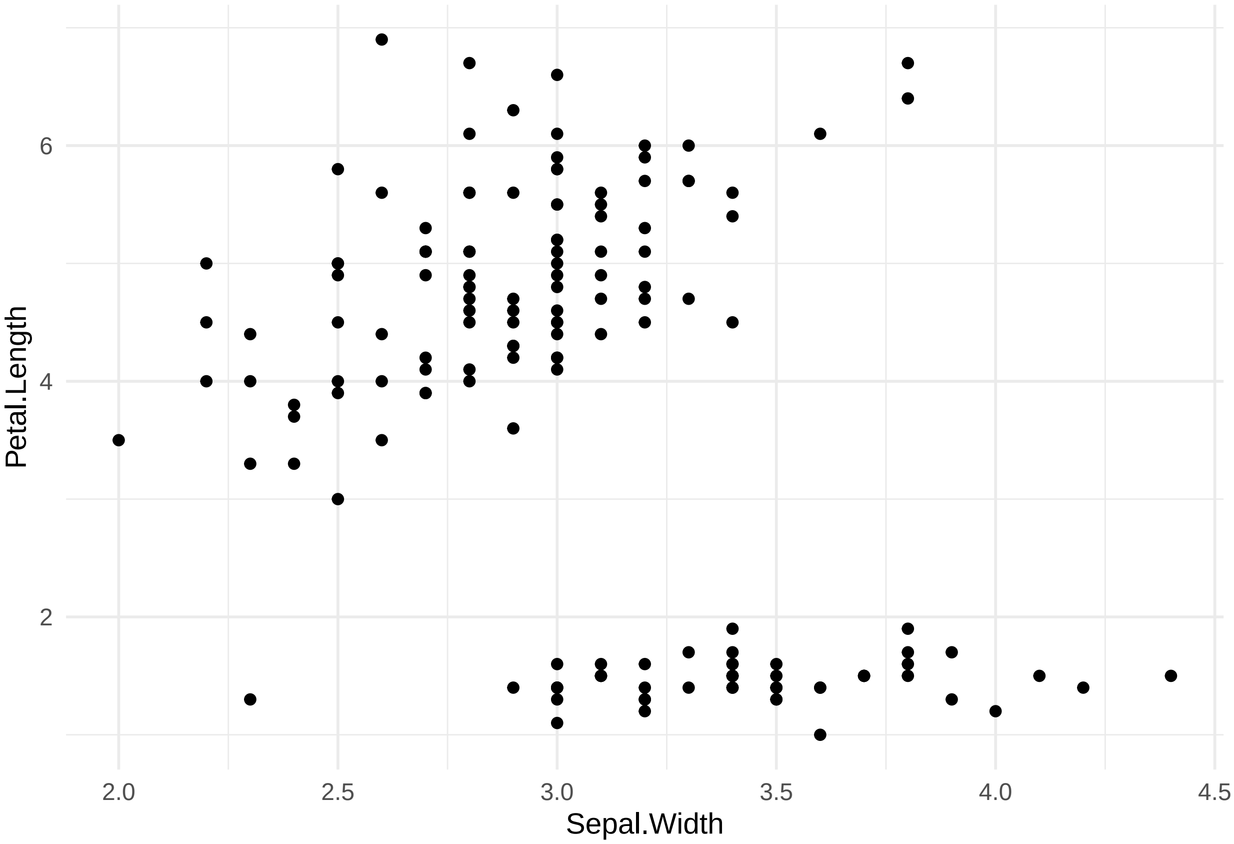
Exercise 5

Returning to previous joint model,

$$\text{petal length} = \beta_0 + \beta_{\text{sepal width}} \cdot \text{sepal width} + \beta_{\text{sepal length}} \cdot \text{sepal length}$$

Does our data offer sufficient evidence that `Sepal.Width` actually helps predict `Petal.Length`?

```
iris %>%
  ggplot(aes(x = Sepal.Width, y = Petal.Length)) +
  geom_point() +
  theme_minimal()
```



To answer this, let's conduct a hypothesis test in a regression framework to find out.

If `Sepal.Width` does not predict predict `Petal.Length`, $\beta_{\text{sepal width}} = 0$, this is our null hypothesis.

- What is the alternative?

H_A: $\beta_{\text{sepal width}} \neq 0$

For OLS regression, our test statistic is

$$T = \frac{\hat{\beta} - 0}{\text{SE}_{\hat{\beta}}} \sim t_{n-2}$$

We want to see if our observed statistic, $\hat{T}$, falls far in the tail under the null.

`R` takes care of much of this behind the scenes with the tidy output and reports a p-value for each β by default.

Fit the regression model and display the tidy output below.

```
# code here
fit3 %>%
  tidy()
```

```
## # A tibble: 3 x 5
##   term         estimate std.error statistic  p.value
##   <chr>           <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)     -2.52    0.563      -4.48 1.48e- 5
## 2 Sepal.Length     1.78    0.0644     27.6  5.85e-60
## 3 Sepal.Width     -1.34    0.122     -10.9  9.43e-21
```

Is $\beta_{\text{sepal width}}$ significant?

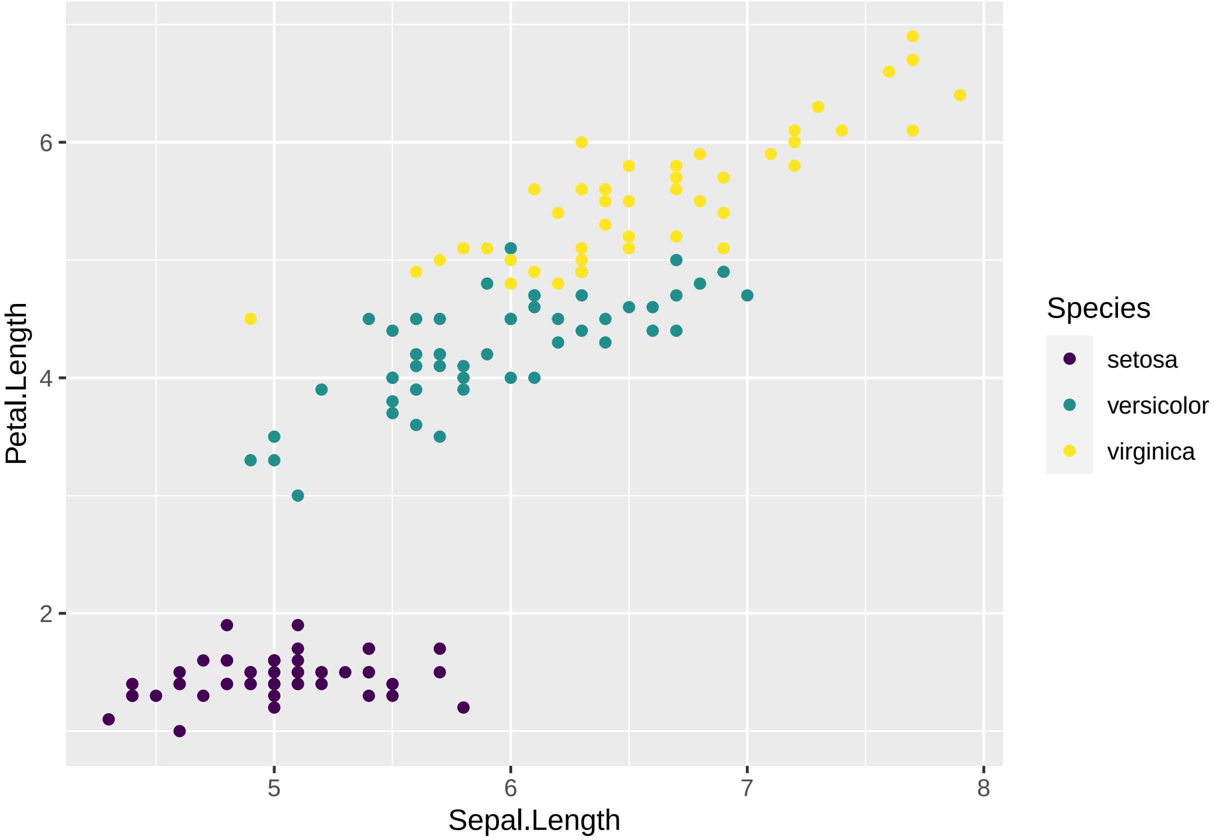
$\beta_{\text{sepal width}}$ is statistically significant with a p-value less than 5.8×10^{-21}, wow!

Also we can calculate this by hand above using `pt`.

Interactions

Exercise 6

```
iris %>%
  ggplot(aes(x = Sepal.Length, y = Petal.Length, color = Species)) +
  geom_point() +
  scale_color_viridis_d()
```



In the plot above, it appears the relationship between sepal length and petal length, i.e. the slope of `Petal.Length ~ Sepal.Length` varies drastically from one species of iris to another.

- Fit a model with predictors `Sepal.Length`, `Species` and an interaction effect between `Sepal.Length` and `Species`. See here for example from the prep.

```
# code here
linear_reg() %>%
  set_engine("lm") %>%
  fit(Petal.Length ~ Sepal.Length * Species, data = iris) %>%
  tidy()
```

```
## # A tibble: 6 x 5
##   term                              estimate std.error statistic    p.value
##   <chr>                                <dbl>     <dbl>     <dbl>      <dbl>
## 1 (Intercept)                        0.803     0.531     1.51  0.133
## 2 Sepal.Length                       0.132     0.106     1.24  0.216
## 3 Speciesversicolor                 -0.618     0.684    -0.904 0.368
## 4 Speciesvirginica                  -0.193     0.658    -0.293 0.770
## 5 Sepal.Length:Speciesversicolor     0.555     0.128     4.33  0.0000278
## 6 Sepal.Length:Speciesvirginica      0.618     0.121     5.11  0.00000100
```

- Write the fitted linear model below (using x, y notation) but replacing βs with their fitted values.

$$y = 0.8 + 0.1 \cdot x_1 - 0.6x_2 - .2x_3 + .55x_1x_2 + .61x_1x_3$$

y: petal length

x_1: sepal length

x_2: is versicolor species? 1 if yes, 0 if not.

x_3: is virginica species? 1 if yes, 0 if not.